PROCEEDINGS OF TDWG

Biodiversity Information Science and Standards 1: e20339
doi: 10.3897/tdwgproceedings.1.20339

OPEN ACCESS

Conference Abstract

BHL: A Source for Big Data Analysis

Mike Lichtenberg ‡

‡ Biodiversity Heritage Library / Missouri Botanical Garden, St. Louis, United States of America

Corresponding author: Mike Lichtenberg (mike.lichtenberg@mobot.org)

Received: 15 Aug 2017 | Published: 16 Aug 2017

Citation: Lichtenberg M (2017) BHL: A Source for Big Data Analysis. Proceedings of TDWG 1: e20339.
https://doi.org/10.3897/tdwgproceedings.1.20339

Abstract

The Biodiversity Heritage Library (BHL) is a consortium of natural history and botanical libraries that cooperate to digitize taxonomic literature and to make that literature available to a global audience for open access and responsible use as a part of a global “biodiversity commons”. In partnership with the Internet Archive and through local digitization efforts, BHL has digitized more than 200,000 volumes of taxonomic literature. Using the Global Names Recognition and Discovery (GNRD) service, BHL has identified over 177 million instances of species names (including more than 29 million unique names) within the text. This content, which includes over 52 million pages of text, provides a rich unstructured source of biodiversity big data that is associated with taxonomic and bibliographic metadata. BHL allows users to search the collection, read the texts online, and download select pages or entire volumes as PDF files. More importantly, BHL makes the source data available for reuse and Big Data analysis via a number of different services. These services include direct downloads of data files and machine interfaces. This talk will describe the downloads and machine interfaces through which BHL source data is available. Each service will be introduced and described. Where feasible, the usage of each service will be demonstrated. Theoretical examples of how each service could be used to facilitate Big Data analysis will be provided.

Keywords

BHL, Big-Data, OCR, Metadata

Presenting author

Mike Lichtenberg